SRI International

Technical Report—Year 1 • March 1994
Covering the Period 4 August 1992 to 4 August 1993

HIGH-PERFORMANCE SPEECH RECOGNITION USING CONSISTENCY MODELING

Vassilios Digilakis, Research Engineer
Peter Monaco, Research Engineer
Hy Murveit, Principal Engineer
Mitchel Weintraub, Senior Research Engineer

Speech Technology and Research Laboratory

SRI Project 3773
Contract N00014-92-0154
Effective 4 August 1992 to 4 August 1994

Prepared for:

Office of Naval Research
Ballston Tower One
800 North Quincy Street
Arlington, VA 22217-5000

Attention: Lt. Col. Robert Powell
Engineering Sciences Directorate

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Projects Agency of the U.S. Govenment.

1. TECHNICAL SUMMARY

The goal of this project conducted by SRI International (SRI) is to develop *consistency modeling* technology. Consistency modeling aims to reduce the number of improper independence assumptions used in traditional speech-recognition algorithms so that the resulting speech-recognition hypotheses are more self-consistent and, therefore, more accurate. Consistency is achieved by conditioning HMM output distributions on state and observations histories, $P(x/s,H)$. The technical objective of the project is to find the proper form of the probability distribution P, the proper history vector, H, and the proper feature vector, x, and to develop the infrastructure (e.g. efficient estimation and search techniques) so that consistency modeling can be effectively used.

During the first year of this effort, SRI focused on developing the appropriate base technologies for consistency modeling. We developed *genonic* hidden Markov model (HMM) technology, our choice for P above, and *Progressive Search* technology for HMM systems which allows us to develop and use complex HMM formulations in an efficient manner. Papers describing these two techniques are included in the Appendix of this report, and are briefly summarized below. This report also describes other accomplishments of Year 1, including the initial exploitation of discrete and continuous consistency modeling and the development of a scheme for efficiently computing Gaussian probabilities.

All of our work that is aimed at improving the word accuracy of speech recognition systems is regularly evaluated using standard test sets. Table 1 shows our progress in terms of reduced word error rate. Our overall error rate was reduced by 41%.

We performed two studies aimed at removing independence assumptions from HMM systems. Our initial attempt, local discrete-density HMMs, did not provide us with an improvement in accuracy. A subsequent study, local continuous-density consistency modeling, made clear some of the problems with the initial approach (e.g. poor choices of context and output distribution type) and shows great promise for future reductions in speech recognition error rates.

2. SPEECH RECOGNITION ACCURACY IMPROVEMENTS

Table 1 shows the reduction in error rate we achieved on ARPA's November 1992 *Wall Street Journal* (WSJ0, 5K bigram test). We evaluated SRI's Decipher[1] technology as it existed at the start of this project in ARPA's November 1992 evaluation. It was a tied-mixture HMM using SRI's phone set and a combination of SRI's and Dragon's WSJ pronunciation dictionaries. We achieved a 13% error rate. In June 1993, after improving our system and conducting regular progress checks with other development materials (using different speakers than the November 1992 test set), we reevaluated our speech recognition system on the November 1992 test set.[2]

1. Decipher is a trademark of SRI International.

Table 1 shows that improvements made on the choice of phonetic units, the dictionary (supplied

Table 1: Speech Recognition Accuracy Improvement

System	Word Error (%)	Sentence Error (%)
SRI Nov. 1992	13.0	73.9
PTM + Cepstral Mean Removal + Phone-set + Dictionary	9.0	60.6
Genones + above improvements	7.7	53.0

by LIMSI), a cepstral mean removal front-end, and, in particular, the use of phonetically tied mixtures (PTM, see the Appendix) reduced our system's error rate by 31%.[3]. An additional 14% reduction was achieved by the introduction of genone technology, making the overall improvement 41%. We expect that genone technology will be even more effective with the increased amount of training data available in Fall 1993 (WSJ1).

3. GENONE-BASED HMM TECHNOLOGY

SRI has developed a new type of hidden Markov model speech recognition technique called genonic mixtures, or genones. In this type of system, Gaussian mixture components are shared among groups of states. These groupings are automatically determined using agglomerative clustering techniques. This technique automatically balances the modeling resolution/robustness trade-off depending on the amount of training data. As we stated in the previous section, by using this and other techniques, SRI has reduced its word error rate on ARPA's November 1992 baseline 5,000 word *Wall Street Journal* bigram evaluation set from 13.0% to 7.7%, a change of 41%.[4]

The genone technique is important for consistency modeling because we plan to base our consistency modeling systems on conditional Gaussian output distributions. Because of limited training data, these high dimensional distributions will require us to use parameter smoothing to maintain a careful balance between high resolution and robustness. The genone technique should permit us to achieve the best performance possible given the training data available. A paper describing this technique has been included in the Appendix.

4. PROGRESSIVE-SEARCH TECHNOLOGY

Another technique called Progressive Search has been developed that allows recognition experiments to be run over several hundred sentences in a few hours instead of a day or more.

---

2. The November 1992 test set was only used twice, once in November 1992 and the second time in June 1993. Because the particular errors made in November 1992 were not examined, we consider this second test to be a relatively fair evaluation of our progress.
3. Our development data experiments suggest that about one half of the 31% improvement is due to PTM.
4. An error rate reduction of 25% was due solely to genone technology.

Progressive Search is a multiple-pass technique, with each pass using a progressively more accurate (and costly) algorithm. The output of each pass is a grammar (word lattice) which is used to constrain the next pass's search space (instead of a less efficient N-best sentence list). It allows evaluation of computationally demanding algorithms (N-grams, more complex HMMs). It also facilitates developing real-time high-accuracy large-vocabulary recognition.

A Progressive Search technique has been applied to a standard cross-word tied-mixture 5K bigram HMM recognizer for ARPA's *Wall Street Journal* dictation task. It improved recognition development time by an order of magnitude (from 46 x real time to 5.6 x real time) when precomputed first-pass lattices were stored.

Another important application of the Progressive Search technique has been for trigram language models. In this case, the grammar output by an initial bigram-based recognizer was converted into a trigram grammar by replicating those states in the grammar where trigram word transition probabilities existed. This approach to trigram language modeling increased decoding time only slightly from that of bigram (15% increase), with only a minimal increase to the grammar size (since most of the trigrams were not represented). This approach is much more powerful than using an N-best approach to implementing N-gram language models since more of the correct words exist in the lattice than the top N sentences. For example, in a recent experiment using bigram language models for a 5,000-word *Wall Street Journal* speech recognition system, a system that achieved approximately a 10% word-error rate[5] on our development set achieved only an approximately 5% N-best error rate[6] for N = 1000, whereas the relatively compact grammar generated by this system had a 1% lattice error rate.[7] This reduced error rate gives the language model the opportunity to repair errors that the N-best system could not overcome. A paper describing this technique has been included in the Appendix.

5. EFFICIENT COMPUTATION OF GAUSSIAN PROBABILITIES

Our genonic HMM systems represent about 25,000 Gaussian distributions per feature when implementing a 5,000-word WSJ speech recognition system. In this case, evaluating Gaussian distributions in training and recognition dominates computation times. We have developed a decision-tree-based scheme, similar to a tree vector quantizer, in which Gaussian distributions are evaluated only if there is good reason to believe their probability density is high for the current observations. This technique significantly reduces the number of Gaussian evaluations, and should easily extend to the conditional segmental Gaussian distributions proposed above. This scheme is also very important when reducing computation times in decoding.

---

5. A 10% bigram error on our development set is roughly equivalent to a 7% word error using bigrams on the official November 1992 evaluation set, approximately the same as the best bigram-based performance reported at the January 1993 ARPA meeting.
6. The N-best word error rate is defined as average error of the best of the N sentence hypotheses.
7. The lattice error rate is the average of the error rate associated with the best path through the lattices.

6. LOCAL DISCRETE-DENSITY CONSISTENCY MODELING

Local-consistency modeling attempts to remove the independence assumption of nearby frames, but not frames across the entire input sentence. The spectral input to the HMM system at neighboring frames is highly correlated because (1) the speech signal is sampled faster (every 10 ms) than the vocal tract changes, and (2) the spectral analysis between neighboring frames uses overlapping windows (25.6 ms). Therefore, the HMM independence assumption is clearly violated; the goal is to modify the HMM model to capture this correlation between neighboring frames.

Our goal is to replace the standard output distribution $p(x_t / s_j)$ with a model that can account for the previous acoustic history $p(x_t / H_t, s_j)$, where H_t is the summary of the previous acoustic input. A straightforward implementation is to represent the summary of the previous acoustics H_t by x_{t-1}. The current frame is highly correlated with the previous acoustic frame, and although prediction of the current frame using a longer history and spectral dynamics is theoretically better, a good first-order approximation that uses only the last observation may be sufficient.

In a discrete density system, the goal is therefore to compute the state conditional output probabilities shown in Eq. (1):

$$p(q_t \mid q_{t-1}, s_j) \tag{1}$$

where q_t is the vector-quantized speech signal at time t. To compute this quantity directly would be extremely difficult and would require the estimation of a very large number of parameters ((num_states = 10,000) x (codebook_size = 256)^2 = 650 million parameters per feature). To reduce the number of parameters, we need some type of model of the relationship between parameters at neighboring frames. The model we have developed is shown in Eq. (2):

$$p(q_t \mid q_{t-1}, s_j) = p(q_t \mid s_j) \times \frac{p(q_t \mid q_{t-1}, s_j)}{p(q_t \mid s_j)} \tag{2}$$

We can approximate the likelihood ratio denoted by the second term in Equation 2 by replacing the context dependent state s_j with its corresponding context independent state c_i. The approximation of the conditional distributions can be computed as:

$$p(q_t \mid q_{t-1}, s_j) \approx \frac{p(q_t \mid s_j) \times \frac{p(q_t \mid q_{t-1}, c_i)}{p(q_t \mid c_i)}}{\sum_{q_t} p(q_t \mid s_j) \times \frac{p(q_t \mid q_{t-1}, c_i)}{p(q_t \mid c_i)}} \tag{3}$$

The above approximation reduces the number of parameters that need to be computed dramatically ((num_states = 180) x (codebook_size = 256)^2= 12 Million parameters per feature). The number of states is 180 because there are 60 context-independent phones with 3 states each. Each context dependent state uses the local-dependency information from its corresponding context-independent state. This approach was tested on one of our *Wall Street Journal* development test sets; the results are summarized in Table 2.

Table 2: Word Error for WSJ Male 5K Closed Verbalized Punctuation Development Test

Model	Standard Recognizer	Recognizer with Co-Occurrence Local Consistency
Context-Independent	46.2	41.8
Context-Dependent	20.7	22.0

While the context-independent model results improved, the context-dependent model performance worsened. We believe that this result is due to the poor estimation of the likelihood ratio parameters because of the large number of parameters (12 million/feature). The number of parameters increases proportionately with the square of the codebook size. It is, therefore, essential to reduce the codebook size, and this can be achieved with phonetically tied mixtures or genones (described in the Appendix). We estimate that this will allow us to reduce the number of parameters by an order of magnitude—reduction from 256 VQ probabilities to 50 Gaussian mixture weights will reduce the number of parameters by a factor of 25.This approach can be combined with our continuous-distribution approach described in the next section.

7. LOCAL CONTINUOUS-DENSITY CONSISTENCY MODELING

For a given HMM state sequence, the observed features at nearby frames are highly correlated. Modeling time correlation can significantly improve speech recognition performance for two reasons. First, dynamic information is very important [Furui86], and explicit time-correlation modeling can potentially outperform more traditional and simplistic approaches like the incorporation of cepstral derivatives as additional feature streams. Second, sources of variability such as microphone, vocal tract shape, speaker dialect, and speech rate will not dominate the likelihood computation during Viterbi decoding by being rescored at every frame.

The output-independence assumption is not necessary for the development of the HMM recognition (Viterbi) and training (Baum-Welch) algorithms. Both of these algorithms can be modified to cover the case when the features depend not only on the current HMM state, but also on features at previous frames [Wellekens87]. However, with the exception of our earlier work [Digalakis93a] that was based on segment models, explicit time-correlation modeling has not improved the performance of HMM-based speech recognizers [Brown87, Kenny90].

To investigate these results, SRI conducted a study to estimate the potential improvement in recognition performance when using explicit correlation modeling over more traditional methods like time-derivative information. We used information-theoretic criteria and measured the amount of mutual information between the current HMM state and the cepstral coefficients at

a previous "history" frame. The mutual information was always conditioned on the identity of the left phone, and was measured under three different conditions:

- *I(h,s)*—unconditional mutual information between the current HMM state and a cepstral coefficient at the history frame; a single, left-context-dependent Gaussian distribution for the cepstral coefficient at the history frame was hypothesized,
- *I(h,s/c)*—conditional mutual information between the current HMM state and a cepstral coefficient at the history frame when the same cepstral coefficient of the current frame is given; a left-context-dependent, joint Gaussian distribution for the cepstral coefficients at the current and the history frames was hypothesized,
- *I(h,s/c,d)*—same as above, but conditioned on both the cepstral coefficient and its corresponding derivative at the current frame.

The results are summarized in Table 3 for history frames with lags of 1, 2, and 4 and a variable one. In the variable case, we condition the mutual information on features extracted at the last frame of the previous HMM state, as located by a forced Viterbi alignment. We can see from Table 3 that in the unconditional case, the cepstral coefficients at frames closer to the current one provide more information about the identity of the current phone. However, the amount of additional information that these coefficients provide when the knowledge of the current cepstra and their derivatives is taken into account is smaller. In addition, the additional information in this case is larger for lags greater than 1, and is maximum for the variable lag.

Table 3: Mutual information (in bits) between HMM state *s* at time *t* and cepstral coefficient *h* at time *t-d* for various lags. Included is the conditional mutual information when the corresponding cepstral coefficient and its derivative at time *t* are given.

Information Lag *d*	0	1	2	4	Variable
I(*h*, *s*)	0.28	0.27	0.25	0.19	0.25
I(*h*, *s* | *c*)	0	0.13	0.15	0.15	0.21
I(*h*, *s* | *c*, *d*)	0	0.11	0.14	0.13	0.20

Our results would predict that the previous frame's observation is not the optimal frame to use when conditioning a state's output distribution. To verify this, and to actually evaluate recognition performance, we incorporated time-correlation modeling in SRI's most accurate recognition system that uses genonic mixtures [Digalakis93b]. The tying of Gaussian mixtures across different HMM states in that system is determined automatically using clustering procedures, and its recognition performance evaluated on the official November 92 *Wall Street Journal* evaluation set is comparable to that of the best reported results. Specifically, we generalized the Gaussian mixtures to mixtures of conditional Gaussians, with the current cepstral coefficient conditioned on the corresponding cepstral coefficient of the history frame. We either replaced the original unconditional distributions of the cepstral coefficients and their derivatives with the conditional Gaussian distributions, or we used them as additional output distributions. The results are summarized in Table 4 for fixed-lag history frames. We can see that the recognition results are in perfect agreement with the behavior predicted by the mutual-

information study. The improvements in recognition performance over the system that does not use conditional distributions are actually proportional to the measured amount of conditional mutual information at the various history frames. However, these improvements are moderate and indicate that the derivative features model the local dynamics effectively. According to the mutual information results, we should expect a significant improvement in recognition performance when modeling the dependencies between the current frame and the last frame of the previous state, i.e., when we model the dynamics across the whole subphonetic segment. Thus, modeling segmental dependencies effectively requires conditioning the output HMM distributions not only on the previous output frames, but also on the segment start time [Ostendorf89].

Table 4: Recognition rates on WSJ corpus with conditional distributions replacing the unconditional ones or used in parallel

Delay	Conditional Only Word Error (%)	Parallel Use Word Error (%)	$I(h, s \mid c, d)$
0	10.32	-	0
1	10.98	10.19	0.11
2	10.50	9.65	0.14
4	10.32	9.83	0.13

The observation that modeling the segment dynamics can improve recognition performance is consistent with previous results by other researchers. The improvement in phone-recognition performance that we previously reported [Digalakis93a] was based on a model that captured the temporal dependencies across the whole phonetic segment through a computationally expensive dynamical system formalism. State-of-the-art recognition performance on the Texas Instruments isolated word database has recently been reported by a dynamic-time-warp-based system that uses segmental features to model the segment dynamics [Algazi93]. We believe that the incorporation of segmental features and the modeling of segment dynamics can significantly improve large vocabulary recognition performance.

7.1 Other Accomplishments

Other accomplishments during the reporting period are described below.

Reducing time required to train HMM systems. SRI's software currently supports computing and storing probabilistic HMM state alignments as computed by the forward-backward algorithm during training. We have found that when we invent new algorithms, we can use alignments that were previously generated by our best system instead of recomputing the alignments with the new algorithms. This saves computation time. Once significant algorithmic improvements have been made, then new alignments should be computed and stored. This approach has reduced our experiment turnaround time by a factor of 2 to 3.

Software infrastructure. We have improved SRI's Decipher speech recognition software by replacing the software that dealt with hidden Markov model state-output distributions with a much more modular software package. This software facilitates experimentation with new state-

output distributions, and, therefore, will be an important tool in developing consistency models for speech recognition. The software package has been completed and is now installed in the Decipher system.

We have also implemented a software package for clustering output distributions. This tool is currently being used to reduce the parameters in our system by tying different states to the same output distribution. This is very important in consistency modeling, since the models that we are investigating have a significantly larger number of parameters than conventional ones. The same tool is also used to define groups of output distributions that share the same Gaussian mixture, as explained in the Appendix.

8. LISTS OF PUBLICATIONS, PRESENTATIONS AND REPORTS

8.1 Refereed papers submitted but not yet published

Digalakis, V. and H. Murveit, "An Algorithm for Optimizing the Degree of Mixture-Tying in a Large-Vocabulary HMM-Based Speech Recognizer," submitted to the IEEE International Conference on Acoustics, Speech and Signal Processing, 1994.

L. Neumeyer, V. Digalakis, M. Weintraub, "Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus," to appear in *IEEE Trans. Speech and Audio Processing*, Special Issue, Spring 1994.

8.2 Refereed papers published[8]

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's Decipher Speech Recognition System: Progressive-Search T echniques," *Proceedings ICASSP-93*.

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Progressive Search for Large Vocabulary Speech Recognition," *Proceedings of the ARPA Human Language Technology Workshop*, March 1993.

8.3 Invited presentations

H. Murveit, "Progressive Search Techniques," ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Massachusetts.

M. Weintraub, "SRI's Stress-Test Benchmark," ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Demonstration of a 20,000-word continuous speech recognition in ARPA's *Wall Street Journal* domain, ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Massachusetts.

---

8. Actually, these papers had extended abstracts that were refereed; the papers themselves were not refereed.

M. Cohen, V. Digalakis, H. Murveit, P. Price, M. Weintraub, "Speech Recognition: an Overview, Examples and Demonstration," presented at Information Systems Laboratory, Stanford University, February 1993.

H. Murveit, Overview and summary talk for the demonstration session of the ARPA Human Language Technology Workshop, March 22, 1993, Plainsboro, New Jersey.

M. Weintraub, "Progressive Search for Large Vocabulary Speech Recognition," ARPA Human Language Technology Workshop, March 22, 1993, Plainsboro, New Jersey.

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's Decipher™ Speech Recognition System: Progressive-Search Techniques," ICASSP-93, April 1993.

V. Digalakis, "Search and Modeling Issues in Large-Vocabulary Speech Recognition" presented at Xerox PARC, August 1993.

9. TRANSITIONS AND DoD INTERACTIONS

We were active participants in the 6-week Robust Speech Processing workshop sponsored by NSA at Rutgers in July-August 1993. Our two researchers there, Leo Neumeyer and Vassilios Digalakis, focused on training issues and channel equalization techniques for acoustic modeling of telephone speech. Their work at the workshop will be included in a special issue of the *IEEE Transactions on Speech and Audio Processing* scheduled for publication in spring 1994.

SRI's Decipher speech recognition technology is being transitioned to Boston University for joint research funded by NSF and ARPA, and we are currently arranging to modify our Decipher technology on SRI internal funds so that ARPA-sponsored research at the Center for Aids Industrial Productivity in collaboration with the David Sarnoff Research Center can take advantage of this technology for research on robust front-end signal processing. In addition, we are discussing with Nancy Chinchor at SAIC, the possibility of using Decipher's technology in work to be conducted for ARPA and for NASA.

Several applications based on Decipher technology were demonstrated at Spoken Language Technology Applications Day last April. This event was attended by over 300 people, about equally divided among government and commercial representatives. Our participation in this event was sponsored by internal funds.

To further the transfer of Decipher technology, SRI has invested significant internal resources toward the development of robust, portable speech recognition software and tools for its use. Several commercial clients are using the resultant technology in their own research or in field trials.

10. SOFTWARE AND HARDWARE PROTOTYPES

The algorithms and software that are developed in this project will be incorporated into the Decipher speech recognition system. We are attempting to commercialize speech recognition based on Decipher and based on tools and other extensions to it that were funded by SRI's IR&D

support. SRI currently has several commercial clients that are in the process of evaluating speech recognition products based on Decipher.

11. APPENDIX

"An Algorithm for Optimizing the Degree of Mixture-Tying in a Large-Vocabulary HMM-Based Speech Recognizer," submitted to the *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994.

"Large-Vocabulary Dictation Using SRI's Decipher Speech Recognition System: Progressive-Search Techniques," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993.